Early Detection of Transition to Multiple Organ Dysfunction Syndrome Using Physiological Time Series Data

by

Michelle Chyn

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Master of Science and Engineering

Baltimore, Maryland

October, 2018

Abstract

Multiple organ dysfunction syndrome (MODS) has an incidence rate of between 11 to 56% in the PICU. Early prevention and treatment of MODS is important in the pediatric population as it increases mortality and leads to possible negative functional outcomes in adulthood. MODS severity is measured using a few different metrics, among which the Pediatric Logistic Organ Dysfunction 2 Score (PELOD-2) is the most recent, pediatric multi-center validated scoring system. This study attempted to build a generalized linear model to detect risk of PICU patients at Johns Hopkins Children's Center from a retrospectively gathered cohort, using PELOD-2 Score≥6 to define MODS severity and minute to minute physiological data as model covariates. Patient specific models were built with a two hour window for transitioning into severe state, the positive class, and the non-severe state was undersampled to balance classes. A global model was built across the majority of the patient population with similar parameters in order to create a more useful, clinical applicable model. The accuracy, sensitivity, and specificity of training and testing sets were calculated for each model. Patient specific models performed well, but performance decayed for the global model, where predictions at the patient level for risk of transitioning had high sensitivity and very low specificity. Future research should continue to refine the definition

of a severe state of MODS and calibrate the sampling scheme with regards to ratio of data points labeled as healthy versus at risk in order to improve global model performance.

Thesis Committee

Sridevi Sarma (Primary Advisor)
Associate Professor
Department of Biomedical Engineering
Johns Hopkins Whiting School of Engineering

Raimond Winslow
Raj and Neera Singh Professor
Department of Biomedical Engineering
Johns Hopkins Whiting School of Engineering

Melania Bembea
Associate Professor
Department of Anesthesiology and Critical Care Medicine - Pediatrics
Johns Hopkins School of Medicine

Acknowledgments

I would first like to express my sincere gratitude towards my adviser Professor Sridevi Sarma, Associate Professor at the Institute of Computational Medicine in the Whiting School of Engineering, for her continuous support of my Master's research. Her patient guidance and capacity for departing her immense knowledge was a tremendous motivation in helping me succeed in researching and writing my thesis.

I would also like to thank my second thesis reader, Professor Raimond Winslow, Raj and Neera Singh Professor and Director of both the Institute for Computational Medicine as well as for my Master's program. His door was always open to me for advice on both my research as well as for my overall academic success as a Master's student.

I am highly grateful for all the time and countless hours of guidance from Dr. Melania Bembea, Associate Professor of the Department of Anesthesiology and Critical Care in Johns Hopkins Children's Center. Her clinical expertise and passionate participation was invaluable in providing me with research direction and improving my understanding of the research questions.

I would also like to thank Christine Kavanaugh and Allison Leventhal from the Office of Graduate Academic Affairs for their encouragement and advice. I wish to also extend my thanks to the lab technician Stephen Granite for helping manage data and to Dr. Pierre Sacré for his incredibly insightful advice.

Finally, I must profess my profound gratitude for the continuous and unfailing support I received from my family and friends. This accomplishment would have not been possible without them. Thank you.

Michelle Chyn

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Early detection of severity of organ failure is important in the prevention of negative outcomes in critically ill children in the Pediatric Intensive Care Unit (PICU). Multiple organ dysfunction syndrome (MODS) has a reported incidence rate of 11-56% in the PICU [1, 2], and is associated with mortality rates ranging from 11-57% [3, 4]. The prevalence of MODS and its characteristic as a continuum of decaying physiology lend it to be a valuable target for study.

In the pediatric population, the Pediatric Multiple Organ Dysfunction Score (P-MODS), the Pediatric Logistic Organ Dysfunction (PELOD) score, the more recent PELOD-2 score were designed to predict mortality in children with MODS [5, 6, 7, 4]. P-MODS uses lactic acid, PaO_2/FiO_2, Bilirubin, Fibrinogen, and Blood Urea Nitrogen values to determine the point based severity of organ dysfunction in the cardiovascular, respiratory, hepatic, hematologic, and renal systems respectively. The state of each system is represented by the measurement of one variable. The Glasgow Coma Score (GCS) and neurological system was considered to have too many factors affecting availability of reliable measurement due to patients on sedation or neuromuscular blockade drugs, and was excluded from the P-MODS score [6]. All of these scores are normally calculated infrequently in practice, and often only at a 24 hour interval for research purposes.

The PELOD score has been validated at multiple PICUs, and has been shown to be suitable for use in assessing pediatric morbidity due to MODS across multiple hospital centers [4]. The score covers the neurological, cardiovascular, renal, respiratory, hematological, and hepatic systems assigning severity points in each organ system across the variables GCS, pupillary reactions, heart rate, systolic blood pressure, creatinine, PaO_2/FiO_2, $PaCO_2$, mechanical ventilation, white blood cell count, platelets, aspartate transaminase, and prothrombin time. In Leteurtre et al. [8], the PELOD-2 score was developed as an updated version of the original PELOD score that overcame the limitation of non-continuous scoring. The PELOD-2 score also assigned points of severity to each organ system, but calculated probability of mortality based on the sum of severity points across all organ systems. PELOD-2 incorporates mean arterial pressure and lactatemia, similar to the Sequential Organ Failure Assessment (SOFA) in adults by Knaus et al. [9] and P-MODS scores, as opposed to systolic blood pressure and heart rate in the scoring of cardiovascular organ dysfunction. Hepatic dysfunction was left out of PELOD-2 because it contributed very little to variance in performance and prediction of death [8, 10].

It is important to note that these scores were created for use as an outcome measure, and reflect the state of patients at the time that their samples were collected for laboratory testing in relation to probability of mortality. Thus, score rates are captured infrequently as a reflection that several variables are only available after clinician deems that a patient is unhealthy and need blood sampling for laboratory measurements to be taken. The current scoring methods were not designed to predict if a patient was at risk for further progressing into increasingly severe states of organ dysfunction. These pediatric scoring systems were originally based on adult scoring systems for calculating risk scores related to adult mortality, such as the Acute Physiology and Chronic Health Evaluation (APACHE) II and Sequential Organ Failure Assessment (SOFA) [11, 9, 10].

Predicting mortality may be important in adult populations due to their higher mortality rates (29.1% in [12]) related to MODS than in children (5.0%, 6%, and 18.5% in [13, 4, 14]). (The upper limit to the above reported range of pediatric related MODS mortality of 57% comes from an evaluation of PICU standard of care in Malaysia [3]; in more developed countries, the mortality rate is commonly at least three fold lower.) However, pediatric patients who survive MODS are likely to have negative functional outcomes that affect later quality of life and cause long term impairment [15]. Additionally, the progression of MODS occurs more rapidly in children compared to adults [2, 16, 1, 17], with maximum organ failure occurring within 24 h in almost 80% of patients who develop MODS in the PICU [15]. Jaramillo-Bustamante et al. [18] observed that 45.1% of children were admitted to the PICU already in MODS. This study was conducted in Colombia and admits that patients were probably admitted late because of having to transport patients from more remote areas cities containing specialized care resources. In areas where specialized medical care is more widely accessible, it is therefore important to seek measures that predict the early stages of organ failure and risk of developing MODS itself instead of mortality in PICU populations.

Early treatment has been shown to improve patient outcomes in several studies on sepsis, one of the most common causes of MODS in critically ill patients [19, 14]. It is well known that adequate fluid resuscitation and appropriate administration of antibiotics are associated with septic shock prevention and lower mortality in sepsis patients [14, 20]. MODS and sepsis occur in overlapping patient populations [16, 21, 22] and are both characterized by a continual illness progression [22, 23]. Both syndromes also happen to be scored in severity by several of the same physiological variables. SOFA is calculated from PaO_2/FiO_2, platelet count, GCS, mean arterial pressure, administration of vasopressors, and creatinine. All of these variables are used in either the PELOD-2 score or P-MODS with the exception of administration of vasopressors. The success of early treatment in sepsis may be similarly replicated in patients with MODS if the progression is detected and mitigated earlier as well.

Early detection has been been studied at multiple levels of treatment centers at various stages of syndrome progression. The SOFA score, used in both adult MODS and sepsis, has been shown to be a good predictor of early risk for septic patients in Innocenti et al. [11]. SOFA score was calculated on admission to the emergency department high dependency unit and after 24 hours. The 24 hour score was significantly higher in patients who where subsequently transferred to the ICU. While the AUC value of 0.8 was reported in this study, sensitivity and specificity values were not specifically annotated and were around 75% for both values interpreted from the ROC curve. Ghanem-Zoubi et al. [24] assessed the prognostic value of the Rapid Emergency Medicine Score (REMS) on mortality in adults admitted to the general internal medicine department with sepsis.

REMS is based on measurements of age, heart rate, temperature, mean arterial pressure, respiratory rate, peripheral oxygen saturation, and GCS. Many of these physiological parameters overlap with those important in diagnosing MODS. The aptitude of the REMS score to distinguishing survivors from non-survivors was evaluated through the ROC curve, and revealed an acceptable AUC value of 0.79, sensitivity of around 75% and specificity around 70% (values were read off the ROC curve and not exact). However, both Innocenti et al. [11] and Ghanem-Zoubi et al. [24] studied the adult population, and patients with sepsis instead of MODS.

Sweney et al. [25] observed the predictive performance of the pediatric modified sequential organ failure assessment score (M-SOFA) on anticipating if patients required physician intervention in a PICU population compared to the decision of trained clinicians retrospectively evaluating patients' transport records and hospital medical charts. The M-SOFA modifies SOFA score thresholds across each variable to match pediatric age relevant thresholds. M-SOFA was calculated from the first laboratory measurement, and if no measurement was obtained within 6 hours of admission, they were assumed to be normal. Varying the threshold of M-SOFA score produced a ROC curve where the best sensitivity and specificity did not outperform physician triage. The results of this study assumed that all of the actual decisions that were made regarding patients needing to receive medical intervention were ground truth. A secondary outcome was evaluated by correlating M-SOFA with the pediatric risk of mortality III (PRISM-III) score, a validated mortality predictor [26], which revealed no significant correlation between M-SOFA scores and PRISM-III. M-SOFA did provide a high sensitivity (100%) and specificity (87%) in predicting mortality following cardiac surgery in neonates [27], but this result was based solely on a population where all patients have cardiac related comorbidities.

There exists a gap in studies that predict the risk of negative prognoses that include worsening outcomes that occur before death in the pediatric population. Furthermore, all of the above mentioned methods of prediction rely on measurements taken fixed windows. This disregards the continuous nature of MODS and information from physiology that may be gleaned within a pre-determined window where scoring occurs. One attempt to bridge this gap is the study by Sandri et al. [28], in which investigators built Bayesian Networks to predict the probability of sequential organ failure. The SOFA score was used to identify organ dysfunction across the respiratory, cardiovascular, hepatic, renal, neurological, and hematologic systems in 24 hour time slices. A dynamic Bayesian Network was trained with nodes representing up to three concurrent organ failures. This Bayesian approach revealed probabilities of day by day sequences of organ failures starting from admission into the ICU. Accuracy of predicting the three organ failure nodes was 71.62%, 75.54%, and 74.95%. Although this study does focus on predicting the sequence of organ failures instead of mortality, the time slices were still 24 h, a very long time in the scope of quickly evolving pediatric organ dysfunction. Additionally, the nature of the Bayesian model does not reveal information on the underlying physiology of why the sequence of organ failures has the probability to progress in the way the model results present.

Considering that the current research in MODS has not offered a satisfactory predictor of early risk indicators, and that studies have mostly used large time windows, there exists a need for deeper examination of the features significant in pre-organ failure state. Further elucidation of the physiological variables relating to early organ failure states may also provide targets for timely intervention. This study intends to create a model for risk of developing multiple organ failure in the PICU, using data from the pre-organ failure state. The aim of this thesis is to detect early that a patient at risk of entering into a severe organ failure state will make this transition by: defining acceptable detection criteria for a severe state, building a generalized linear model using physiological time series data based on this criteria, and evaluating the performance of the model across a PICU population.

Chapter 2

Methods

2.1 Data Source

Electronic Health Record (EHR) data was acquired from Allscripts Sunrise/POE, and minute to minute physiological data was acquired from HL7 exports of GE Aware Gateways. Data contained 2711 PICU admissions of patients at Johns Hopkins Children's Center between July 2014 to October 2015. Only the first PICU admission was used in the case of patients with multiple admissions [30]. Demographic information of all first PICU admissions is in Table 2.1.

This study was approved by the Johns Hopkins University School of Medicine Institutional Review Board.

2.2 Patient Selection

The overall patient selection criteria are shown in Figure 2.1. The first two blocks are related to data processing, and the latter blocks refer to patient selection after applying the below methodology.

Variable	All Patients (N=2512)	Survivors to PICU discharge (N=2480)	Non-Survivors to PICU discharge (N=32)	p-value
Length of Stay (Days) Median (IQR,Range)	1.85 (0.98 to 3.94, 0.00 to 151.99)	1.84 (0.98 to 3.93, 0.00 to 151.99)	1.98 (0.51 to 12.26, 0.02 to 92.23)	<.0001
Age at Admit (Days) Median (IQR,Range)	1965.92 (555.21 to 4516.72, 0.73 to 6572.44)	1970.60 (569.09 to 4535.72, 0.73 to 6572.44)	1014.84 (96.86 to 3829.33, 0.77 to 5339.56)	0.0578
$\leq$30 days	95 (3.78%)	89 (3.59%)	6 (18.75%)	<.0001
30 days <age <12 mo	421 (16.76%)	414 (16.69%)	7 (21.88%)	0.4358
12 mo $\leq$ age <12 yr	1334 (53.11%)	1319 (53.19%)	15 (46.88%)	0.4774
12 yr $\leq$ age <18 yr	662 (26.35%)	658 (26.53%)	4 (12.50%)	0.0735
Gender (0=Female)	1371 (54.58%)	1357 (54.72%)	14 (43.75%)	0.2158
Race				
White	1148 (45.70%)	1132 (45.65%)	16 (50.00%)	0.6233
Black/African Am	856 (34.08%)	847 (34.15%)	9 (28.13%)	0.4749
Asian	114 (4.54%)	112 (4.52%)	2 (6.25%)	0.6398
Hawaiian/Pacific Is	3 (0.12%)	3 (0.12%)	0 (0.00%)	0.8440
Am Ind/Alaska Native	4 (0.16%)	4 (0.16%)	0 (0.00%)	0.8202
Multi Racial	12 (0.48%)	12 (0.48%)	0 (0.00%)	0.6934
All Other Races	274 (10.91%)	271 (10.93%)	3 (9.38%)	0.7797
Unknown	96 (3.82%)	94 (3.79%)	2 (6.25%)	0.4710
Decline to Answer	5 (0.20%)	5 (0.20%)	0 (0.00%)	0.7994
Trauma/Drowning	7 (0.28%)	7 (0.28%)	0 (0.00%)	0.7636
Operative Cardiac - Neonatal	2 (0.08%)	2 (0.08%)	0 (0.00%)	0.0539
Operative Cardiac - Pediatric	17 (0.68%)	16 (0.65%)	1 (3.13%)	0.0539
Non-operative Cardiac - Neonatal	57 (2.27%)	51 (2.06%)	6 (18.75%)	0.0006
Non-operative Cardiac - Pediatric	893 (35.55%)	875 (35.28%)	18 (56.25%)	0.0006
Respiratory - Neonatal	125 (4.98%)	120 (4.84%)	5 (15.63%)	0.0053
Respiratory - Pediatric	928 (36.94%)	915 (36.90%)	13 (40.63%)	0.6642
Non-Cardiac - Neonatal	38 (1.51%)	38 (1.53%)	0 (0.00%)	0.0006
Non-Cardiac - Pediatric	1522 (60.59%)	1514 (61.05%)	8 (25.00%)	0.0006
Mechanical Ventilation	649 (25.84%)	622 (25.08%)	27 (84.38%)	<.0001
Non-Invasive Ventilation	940 (37.42%)	918 (37.02%)	22 (68.75%)	0.0002
ECMO	27 (1.07%)	17 (0.69%)	10 (31.25%)	<.0001
Hemofiltration	28 (1.11%)	21 (0.85%)	7 (21.88%)	<.0001
Peritoneal Dialysis	16 (0.64%)	15 (0.60%)	1 (3.13%)	0.0750
RBC Transfusion	294 (11.70%)	284 (11.45%)	10 (31.25%)	0.0005
Plasma Transfusion	7 (0.28%)	6 (0.24%)	1 (3.13%)	0.0021
Platlet Transfusion	65 (2.59%)	58 (2.34%)	7 (21.88%)	<.0001
Pneumonia	368 (14.65%)	360 (14.52%)	8 (25.00%)	0.0957
Ventilator assisted Pneumonia	8 (0.32%)	8 (0.32%)	0 (0.00%)	0.7477
ARDS	495 (19.71%)	476 (19.19%)	19 (59.38%)	<.0001
Sepsis	239 (9.51%)	228 (9.19%)	11 (34.38%)	<.0001
Renal Insufficiency	118 (4.70%)	111 (4.48%)	7 (21.88%)	<.0001
Cerebral Hemorrhage/Ischemia	21 (0.84%)	21 (0.85%)	0 (0.00%)	0.6013
Seizure	120 (4.78%)	117 (4.72%)	3 (9.38%)	0.2199
TBI	81 (3.22%)	79 (3.19%)	2 (6.25%)	0.3297
Cardiac Arrest	67 (2.67%)	49 (1.98%)	18 (56.25%)	<.0001
Neuromuscular ˆ	307 (12.22%)	303 (12.22%)	4 (12.50%)	0.9614
Cardiovascular ˆ	636 (25.32%)	622 (25.08%)	14 (43.75%)	0.0158
Respiratory ˆ	194 (7.72%)	193 (7.78%)	1 (3.13%)	0.3270
Renal ˆ	64 (2.55%)	64 (2.58%)	0 (0.00%)	0.3575
Gastro ˆ	37 (1.47%)	37 (1.49%)	0 (0.00%)	0.4866
Heme/immune ˆ	23 (0.92%)	23 (0.93%)	0 (0.00%)	0.5844
Metabolic ˆ	261 (10.39%)	258 (10.40%)	3 (9.38%)	0.8498
Congenital/Genetic ˆ	430 (17.12%)	427 (17.22%)	3 (9.38%)	0.2420
Malignancy ˆ	213 (8.48%)	212 (8.55%)	1 (3.13%)	0.2740

Table 2.1: **Patient Demographics** Demographic information of the 2512 first PICU admissions after initial patient selection. ˆ denotes categories defined by Feudtner et al. [29]

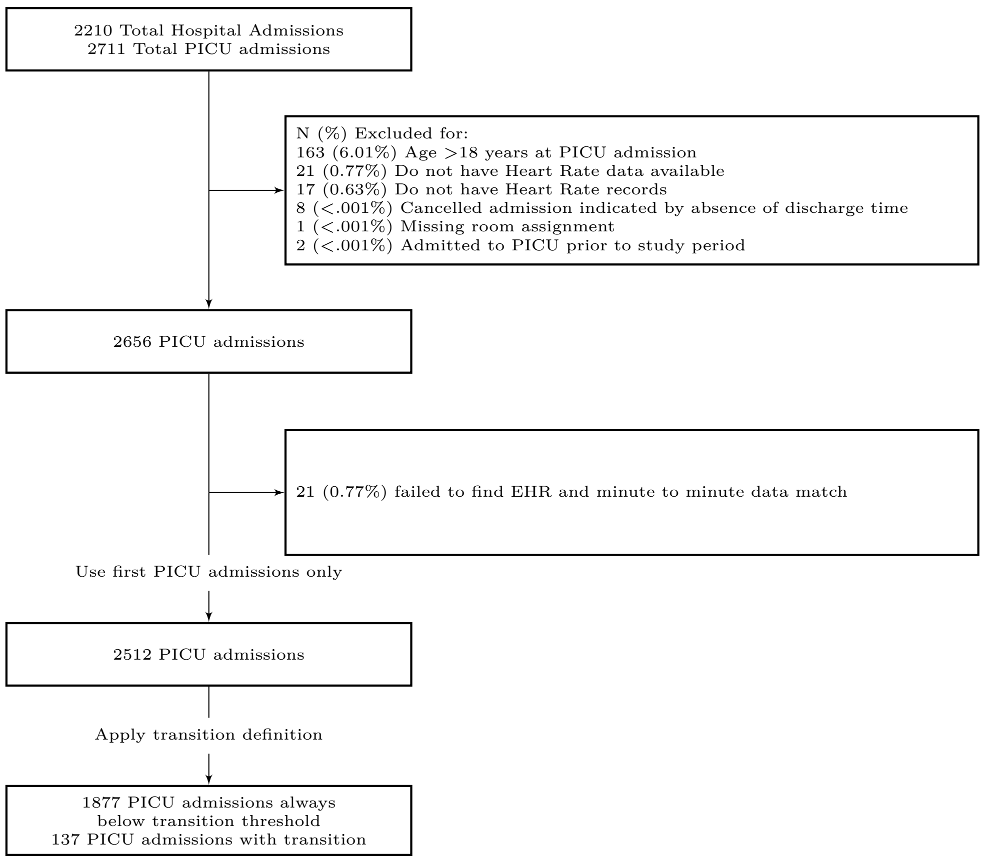


Figure 2.1: **Patient selection process**

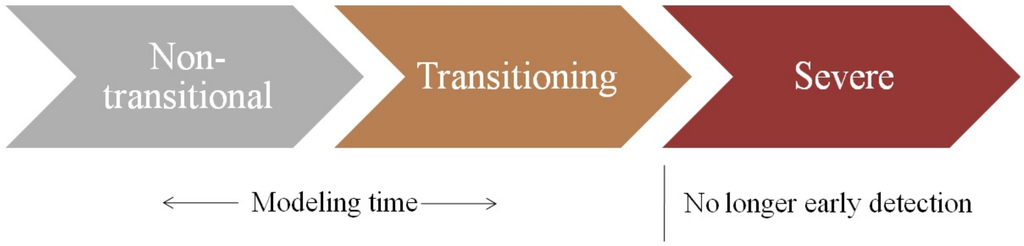


Figure 2.2: **Patient Progression** for a patient who enters severe state

2.3 Patient Severity using PELOD-2 Score

The criteria used for measuring severity of organ failures was the PELOD-2 score based on the Leteurtre et al. [8] definition. However, instead of a daily PELOD score, the PELOD-2 score was calculated in a rolling window. At any time point, the worse measurement in each variable across the previous available time up to 24 hours was used to calculate the score. Missing variables before any value became available were assumed to be normal, adding no points to the score [15]. Values were held for up to 24 hours before they were considered missing again. Further details of data processing EHR data are described in the Appendix 5.2.1.

In order to detect a patient at risk of entering some severe state before the state occured, the model was built using data from patients who are initially healthy before they become unhealthy. The progression of becoming worse was split into two windows of interest: 1) the time where a patient was assumed to be healthy, the non-transitional time and 2) the period before the severe state occured, the transitional window (Fig 2.2).

Here, we start by defining the clinically relevant mark of severe organ dysfunction onset, which was the end of the transitional window. The top portion of Figure 2.3 may be used as an illustrated example of a patient initially relatively healthy, and has a progressively increasing PELOD-2 score. According to the Leteurtre PELOD-2 score, organ dysfunction occurs in each individual organ when the score for variables contributing to one organ is non-zero [8]. This means that a PELOD-2 score of 1 point in Glasgow Coma Score and 1 point in Lactatemia denotes a patient with multiple organ dysfunction, but a patient with PELOD-2 score of 6 points in mean arterial pressure has only single organ dysfunction. In order to create a more continuous scaling of severity, a PELOD-2 score strictly greater than 5 was used to define the transition into a severe clinical state. This score threshold only occurs if a patient either has a very severe score in the cardiovascular associated variable(s), or a combination of two or more organ dysfunctions. In all of the following sections, the crossing of this score threshold will be referred to as the transition, and the time at which this occurs in a patient will be referred to as the transition time.

However, PELOD-2 score alone was not a certain enough to strictly mark a transition. The infrequent capture of variables used to calculate PELOD-2 score created a level of uncertainty in regards to knowing when the transition truly occured. In a sparsely sampled variable such as white blood cell count, a transition could occur anywhere from an hour to some time over 24 hours prior to the recording of a measurement value. After examining the frequency of measurement in each variable contributing to PELOD-2 score (example in Appendix 5.1), only transitions caused by mean arterial pressure and presence of invasive ventilation were counted as trusted transitions, $t = T_i$, time of transition in patient i, for modeling.

Taking into consideration that maximum pediatric MODS typically occurs within the first day of admission [15], that this model was concerned with the period even before maximum severity, and that the median transition time of this study cohort was 2.77 hours after PICU admission, the start of the transitioning window was assigned at 120 minutes before T_i. The maximum number of samples labeled $y_i(t) = 1$ was 120 per patient. It was assumed that the physiology of patients who never crossed the score threshold was in a similar to that of the physiology of the earlier, non-transitional period in patients who did eventually transition. The window of non-transitional time for patients who transition is described in further detail under later sections. Distributions of PELOD-2 Score and organ failure in the patients whose first transitions met the critera described above and those who remained below threshold for the entire PICU stay are shown in Table 2.2.

2.4 Minute to Minute Data Processing

Minute to minute data was collected, containing 6 physiological variables from bedside physiological monitors. For mean arterial blood pressure, systolic blood pressure, and diastolic blood pressure, the non-invasive cuff pressure was used first, and if at any minute cuff pressure was missing, then the arterial pressure was used if available.

Spike outliers were detected across each variable separately using the MATLAB algorithm for median average deviation:

$$\text{Median Absolute Deviation} = \text{median}(|x_i(t) - \text{median}(\mathbf{x_i}(\text{outlier detection window}))|) \tag{2.1}$$

where $x_i(t)$ was a single measurement for patient i at time t and $\mathbf{x_i}$ was the vector of measurements for one variable during a specific outlier detection window. Patient data was split into 2 hour outlier detection windows where the first window began at onset of data, and each subsequent window began 30 minutes after the onset of the previous window to ensure smoothness, until the end of patient data. Values over 3 median average deviations away from the median were counted as outliers. Approximately 10% of patients were randomly chosen to be validated by a clinician to ensure correct window size. Identified outliers were filled using linear interpolation between the previous and next available non-outlier values, and up to 15 minutes of values were replaced.

2.5 Data and Patient Matching

The minute to minute data contained hand entered partial names collected by nurses at the bedside and lacked medical record number (MRN) identifiers, so they were matched to EHR data using room number and time using the following matching rules. It was assumed that one patient would take longer than a minute to switch rooms, so continuous minute to minute recordings were taken as one segment for one patient, and gaps in monitor recording longer than a minute in length were taken as the possible start of a segment belonging to new patient. Each identified segment was matched by endpoints to the segments in the EHR vital signs data, which contained known MRNs. Minute to minute segments with endpoints straddling the room and time of more than one MRN were completely discarded.

Unique manually entered names identified minute to minute segments where multiple streams of data were sometimes recorded in the same room. Although the manually entered names were unreliable, these segments were discarded as we couldn't verify which data stream represented the true patient in the room. Segments where there were no manually entered names were kept, but due to the existence of the multiple stream data problem, they may need further examination.

2.6 Generalized Linear Model Setup

All models constructed from data fell into the class of Generalized Linear Models (GLMs) and in combination with a thresholding rule, classified each patient as having a risk of transitioning into MODS or not. The GLM for a Bernoulli response variable assumes that at any given minute, a patient's state was a Bernoulli random variable, where probability of being in the transitional state was related to a linear combination of patient features[31].

The state of patient i at time t was defined as $y_i(t)$. When $y_i(t) = 1$, at this time t, patient i was in the transitional window before the severe state of interest. $y_i(t) = 0$ signified the opposite, where at time t, patient i was non-transitional. $\mathbf{y}$ was the vector of class labels. For each of these time samples, there was a corresponding feature vector $\mathbf{X_i}(t)$ containing the physiological data at time t.

The probability that patient i at time t was in the transitional state was defined as $p_i(t) = g(\mathbf{X_i}(t))$, where $p_i(t)$ was dependent on a function g operating on $\mathbf{X_i}(t)$, the feature vector of patient i at time t. For a binary outcome regression, g is the logit function, which constrained the probability of positive class outcome $P(y(t) = 1)$ between 0 and 1. The negative class was similarly constrained by definition of being $P(y(t) = 0) = 1 - P(y(t) = 1)$. Thus, the probability that any patient was in the transitional state at time t was defined as

$$P(y(t) = 1) \triangleq p(t) \triangleq g(\mathbf{X}(t), \boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta}^T \mathbf{X}(t)}}{1 + e^{\boldsymbol{\beta}^T \mathbf{X}(t)}} \tag{2.2}$$

Using the maximum-likelihood estimation procedure, the optimal set of coefficients, $\boldsymbol{\beta}$, was found by taking the product of probabilities across time in a designated training sample, and maximizing this likelihood function with respect to $\boldsymbol{\beta}$.

The magnitude and sign of each coefficient resulting from GLM estimation contains potentially clinically meaningful value. For example, since each feature in this model was normalized to a mean of 0 and standard deviation of 1 (described further below), a large positive value coefficient may suggest that the corresponding feature has a large contribution to predicting a patient's transition into MODS.

2.6.1 Patient Specific Modeling

Single patient modeling was used to confirm the rationale behind using a GLM to classify patients transitioning into individualualized severe states of organ dysfunction at an overtrained level. 31 GLMs were built for 31 patients after the selection process and data availability considerations, further discussed in results.

Class Labeling Scheme The transitioning window, began at $t = T_i -$

120 minutes, and all samples before this time were labeled as non-transitional (Fig 2.3).

$$y_i(t) = \begin{cases} 0 & \text{if } t_0 \leq t < T_i - 120 \\ 1 & \text{if } T_i - 120 \leq t < T_i \end{cases} \tag{2.3}$$

where t_0 began at patient admission to PICU or whenever minute to minute data first becames available following PICU admission.

Sampling and Bootstrap For each bootstrapping iteration, each patient was sub-sampled uniformly at random within $\mathbf{y_i} = 0$ and $\mathbf{y_i} = 1$ class labels into 85/15% training and testing data respectively. After being split into training and testing sets, each feature was normalized to a mean of 0 and standard deviation of 1. In the training set, classes were balanced in each patient by further subsampling the class with more samples to the same length as the class with less samples. The MATLAB glmfit function was used to fit the final training set of patients where the resulting number of samples was greater or equal to 90% times number of features plus one (for the constant term), or $\geq$ 63 samples.

$$P(y_i(t) = 1) \triangleq p_i(t) \triangleq g(\mathbf{X_i}(t), \boldsymbol{\beta_i}) = \frac{e^{\boldsymbol{\beta_i}^T \mathbf{X_i}(t)}}{1 + e^{\boldsymbol{\beta_i}^T \mathbf{X_i}(t)}} \tag{2.4}$$

Since this model setup was only being used to substantiate the global population model, only 5 bootstrapping iterations were built per patient.

The 15% of data in each patient was used for testing was not sub-sampled.

Evaluation The criteria for detecting that a patient was at risk of transitioning into severe state per iteration was determined using the decision threshold, τ_i, maximizing sensitivity and specificity on the ROC curve. This is the point on the curve closest to the coordinate $(0, 1)$. The estimated probability $\hat{p}_i(t)$

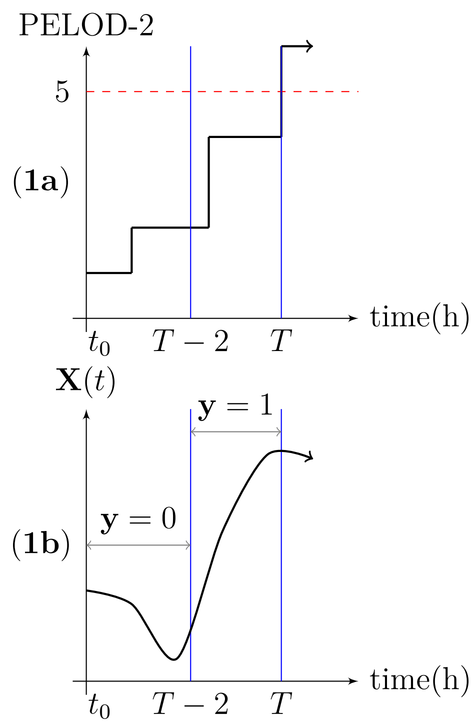


Figure 2.3: **GLM Labels for Patient Specific Model** 1a) Possible PELOD-2 Score of a patient. The blue vertical lines mark the crossing the threshold for transition at $t = T$, and the start of the period labeled as transitioning for the GLM at $t = T - 2$ 1b) Representation of the evolution of an arbitrary feature during the times of interest. The horizontal arrows show the GLM labels in their correct time periods.

was calculated for the training and testing set of each iteration and compared to the decision threshold at each minute.

$$\hat{y}_i(t) = \begin{cases} 0 & \text{if } \hat{p}_i(t) < \tau_i \\ 1 & \text{if } \hat{p}_i(t) \geq \tau_i \end{cases} \tag{2.5}$$

To observe model performance at each iteration, both training and testing sets were evaluated for accuracy, sensitivity, and specificity of their $\mathbf{\hat{y}_i}$ labels across samples. Early detection time, $\hat{T}_i$, was defined as the first time the estimated probability of a patient crossed the decision threshold.

2.6.2 Global Population Modeling

In the model for the global population, patients were pooled together to create one general model. Patients who never transitioned were incorporated as well. In addition to the 6 features from minute-to-minute data used to build the patient specific GLM, indicator features were added to represent age group and gender in the global population model. The constant parameter in MATLAB glmfit was turned off as it was represented in the combination of these indicator features, which are constant per patient.

Labeling Scheme Let patients who contain trustable transitions be denoted as the subset i^1, and patients who never transition to the severe state be denoted as i^0. The class sampling windows used the same definition of a transitional state and the same window length of transitional state as the patient specific model, but incorporated a 120 minute zero-sampling window prior to the transitional state to account for possible differences in the unknown physiology during the very early transitioning time or possible instability in patient physiology after PICU admission. The no-sampling window spanned from $t = T_{i^1} - 240$ minutes to $t = T_{i^1} - 120$ minutes. The non-transitional time began from first available data following PICU admission to 240 minutes prior to T_{i^1}.

Thus, patients who transition follow this labeling scheme

$$y_{i^1}(t) = \begin{cases} 0 & \text{if } t_0 \leq t < T_{i^1} - 240 \\ NaN & \text{if } T_{i^1} - 240 \leq t < T_{i^1} - 240 \\ 1 & \text{if } t \leq T_{i^1} \end{cases} \tag{2.6}$$

while patients who never transition are labeled $y_{i^0}(t) = 0 \forall t$. In patients who never transitioned, only those admitted for more than 120 minutes were used.

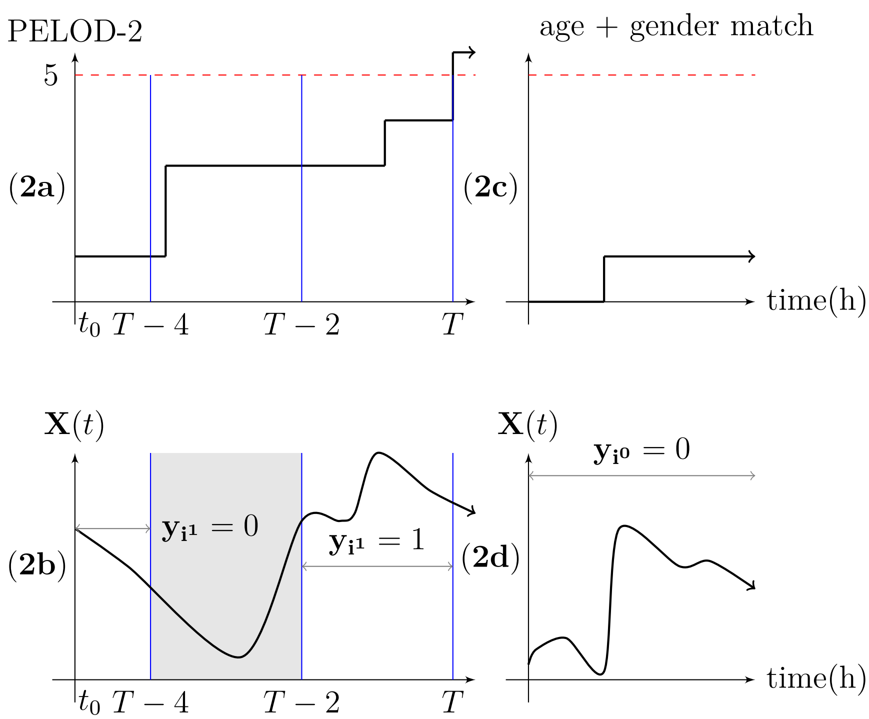


Figure 2.4: **GLM Labels for Global Model** when patient i^1 has sufficient $\mathbf{y_i} = 0$ data available. 2a) Possible PELOD-2 Score of patient i^1 with blue vertical lines marking $t = T$ as the time of transition, $t = T - 2$ hours, and $t = T - 4$ hours the boundaries of the no sampling region, 2b) An example of a model feature overlaid with the GLM labels and a gray box depicting the nosampling region. 2c-d) The possible PELOD-2 Score and single feature of a age group and gender matched i^0 patient.

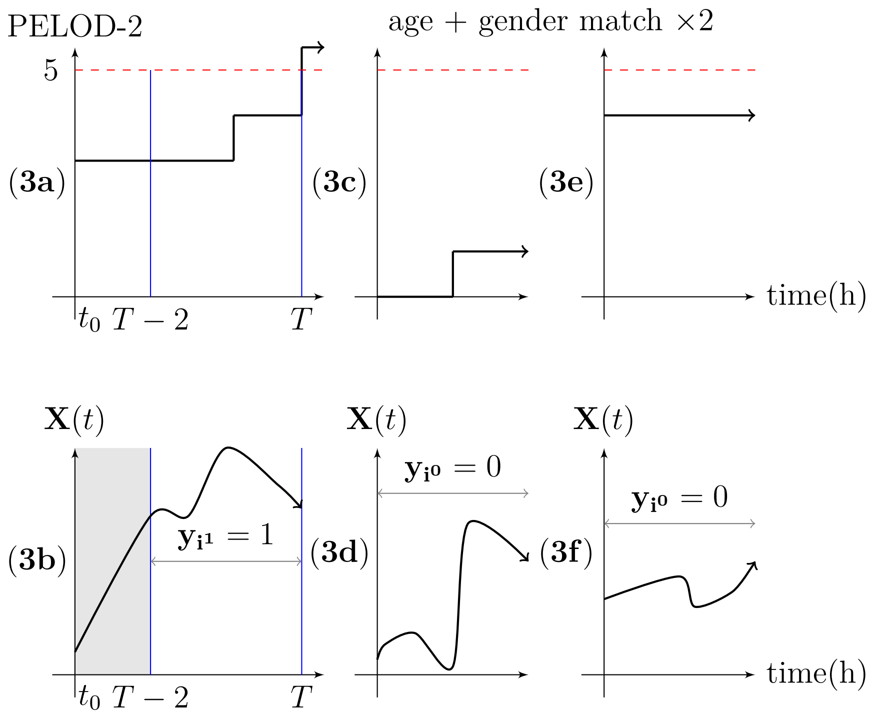


Figure 2.5: **GLM Labels for Global Model** when patient i^1 does not have sufficient $\mathbf{y_{i^1}} = 0$ data available. 3a) Possible PELOD-2 Score of a patient i^1 where they are admitted and undergo first transition within 4 hours, 2b) Example of patient GLM label when there is less than 4 hours of data before transition 3c and e) Example PELOD-2 Scores of age group and gender matched i^0 patients, 3d and f) Example feature values in the two i^0 patients where $\mathbf{y} = 0$ data is sampled from

Bootstrap 50 iterations were used for bootstrapping. In each iteration, 80% of patients containing transitions were use for training, and the remaining 20% for testing. The i^0 patients were split according to the sampling described in the Training and Testing sections below.

Sampling of the Training Set Within the training set of patients, in order to maximize amount of data represented in the model, all transitional $\mathbf{y_{i^1}} = 1$ data was used. Each i^1 patient was gender and age group matched to a i^0 patient from that specific gender and PELOD-2 criteria age group (eg. Male and under 1 month old) uniformly randomly and without replacement from the relevant i^0 patients for one bootstrap iteration. If the length of data in the non-transtional state of patient i^1 was greater than half of the length of transitional data, the $y = 0$ data was randomly sampled and evenly split between patient i^1 and patient i^0, as shown in Figure 2.4. If the length of data in the non-transitional state of patient i^1 was less than the length of transitional data in the same patient, another gender and age group matched i^0 patient was selected, and sampling of non-transitional state data was split evenly between the two i^0 patients (Figure 2.5). This sampling scheme designed in attempt to add more contribution from i^0 patients to the non-transitional state, since the population of i^0 was far larger than the i^1, but still maintain a class balance in number of samples between $y = 0$ and $y = 1$.

Each non-constant feature of the training set was normalized to mean 0 and standard deviation 1 across all patients.

Testing Set The testing set contained the remaining 20% of the i^1 patients and all of the remaining i^0 patients. The testing i^1 patients were labeled with 2 hours of transitional data prior to $t = T_i$, the 2 hour no-sampling window from

	Patients who transition (N = 137)	**Patients always below transition thresold (N = 1877)**	**p-value**
Average PELOD-2 Score in patients Median (IQR, Range)	5.00 (2.00 to 7.00, 0.00 to 21.00)	2.00 (0.00 to 2.00, 0.00 to 5.00)	<.0001
Length of PICU admission in patients (hr)	145.80 (70.20 to 344.90, 0.82 to 2898.00)	35.33 (21.50 to 64.69, 0.00 to 756.33)	<.0001
First transition time (hours after admission)	2.77 (0.45 to 13.36, 0.02 to 958.05)	NaN	NaN
Score at first transition	7.00 (6.00 to 7.00, 6.00 to 8.00)	NaN	NaN
Increase in score amount at transition	3.00 (3.00 to 3.00, 2.00 to 5.00)	NaN	NaN
# occurences of score >0 (during transition or overall):			
Neurological	119	32	<.0001
Cardiovascular	135	1260	<.0001
Renal	44	18	<.0001
Respiratory	137	149	<.0001
Hematalogical	68	157	<.0001

Table 2.2: **Patient Organ Failure Distributions** PELOD-2 Score and transition related distributions in patients containing the desired first transition versus those who always remained below transition threshold.

$t = T_i - 4h$ to $t = T_i - 2h$, and as non-transitional before time $t = T_i - 4h$.

$$y_i(t) = \begin{cases} 0 & \text{if } t_0 \le t < T_i - 4 \cdot 60 \\ NaN & \text{if } T_i - 4 \cdot 60 \le t < T_i - 2 \cdot 60 \\ 1 & \text{if } t \le T_i \end{cases} \tag{2.7}$$

Two normalization procedures were used in the testing set. The first method represented a closer representation to real world application, where patients would be individually normalized to standard normal distributions by each non-constant feature. The second method normalized each non-constant feature across all patients in the testing set.

Evaluation The global population GLM was evaluated in the same way as the patient specific model in Equation 2.5, and sample by sample accuracy, sensitivity, and specifity were calculated for training and testing sets at each iteration. Early detection time was recorded and accuracy, sensitivity, and specificity were calculated across patients to summarize the model performance at detecting on the whole patient level.

Chapter 3

Results

In this chapter, the important study cohort characteristics are summarized, and then the patient specific model results are presented, followed by global model results.

After applying the chosen definition of transitioning into a severe state, where patients were admitted under a PELOD-2 Score of 5 and crossed this threshold score, yielded 5.45% (N=137) PICU admissions for who transitioned and 74.72% (N=1877) who remained below threshold for the entire duration of their PICU stay. The median age of all patients was 5.4 years and ranged from 1 day to 18 years. 54.58% (N=1371) were male. Mortality rate was very low 1.25% (N=32). Other information on comorbidities are listed in Table 2.1.

The median PELOD-2 Score across the entire admissions of patients who transitioned was 5 with a range of 0 to 21. In patients who never transitioned, the median score was 2 with a range of 0 to 5. The median time of first transition was 2.77 hours with an interquartile range of <1 hour to 13.36 hours. This distribution was right tailed, as the longest time of first transition after admission was nearly 18 days. Median length of PICU stay was significantly longer

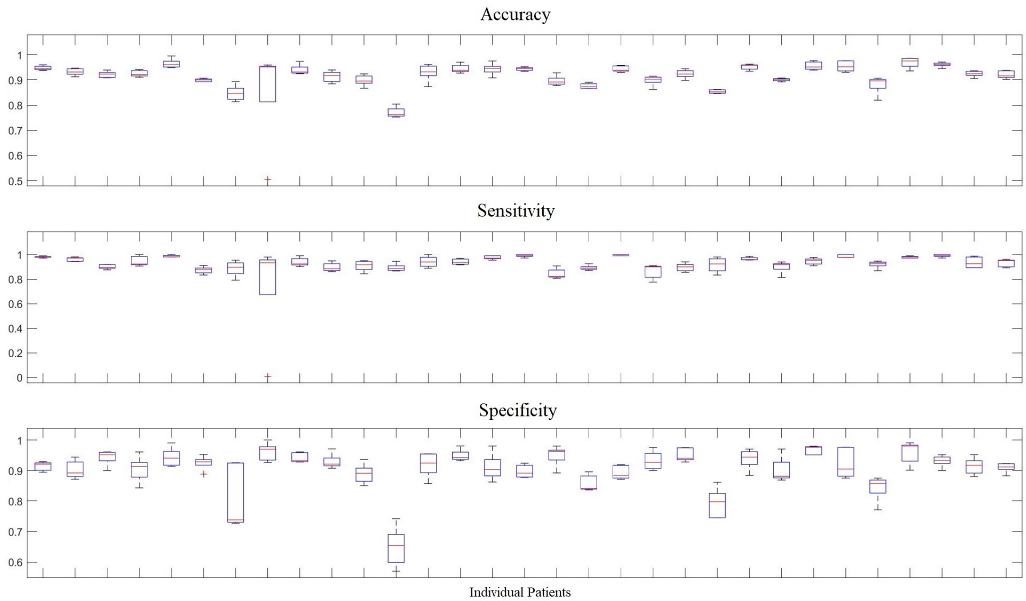


Figure 3.1: **Training Set Statistics** Accuracy, Sensitivity, and Specificity in training sample time point comparisons of $\hat{y}_i(t)$ to $y_i(t)$ across the 5 bootstrap iterations per patient specific model without warnings. The x-axis represent individual patients remaining after sampling and GLM criteria have been met out of the 137 with the desired transition.

in patients who transitioned, at 145.8 hours (∼6 days, p<.05) while patients who remained below transition threshold stayed for a median of 35.33 hours. Individual organ failure presence in patients at transition in the case of those falling within this criteria and presence in entire PICU stay in those who never cross transition threshold are listed in Table 2.2.

3.1 Patient Specific Model

31 Patient admissions remained after patient selection criteria and further omission due to GLM warnings. There were 7 perfect separation warnings, 4 iteration limit warnings, and 2 ill conditioned model warnings. Training model

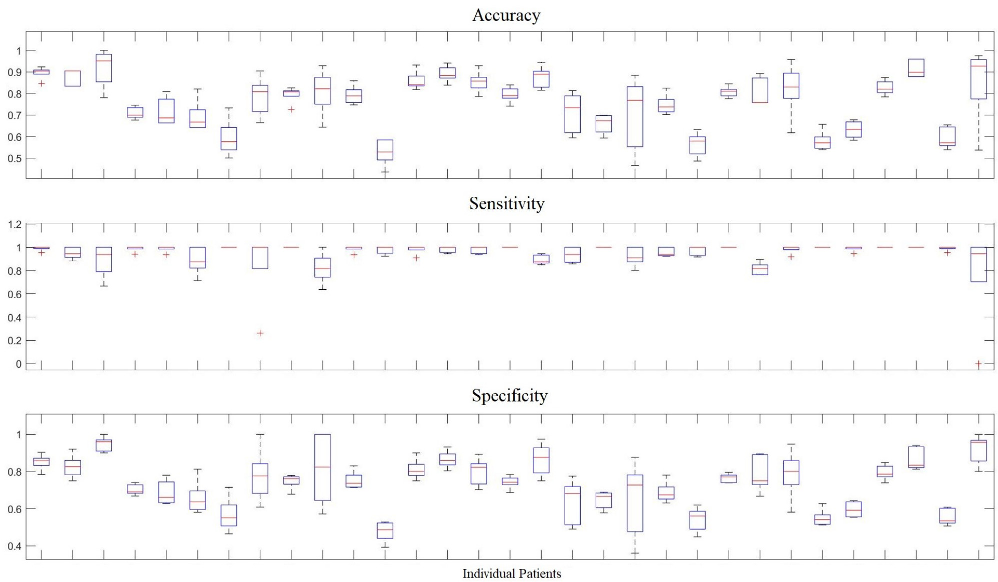


Figure 3.2: **Testing Set Statistics** Accuracy, Sensitivity, and Specificity in testing seample time point comparisons of $\hat{y}_i(t)$ to $y_i(t)$ across the 5 bootstrap iterations per patient specific model without warnings.

performance by patient for comparing model prediction results by time point is depicted in Figure 3.1. Testing performance is shown in Figure 3.2. Across these 31 patient specific models, the training accuracy, sensitivity and specificity was $91.57\% \pm 4.26\%$, $92.65\% \pm 5.28\%$, and $90.49\% \pm 6.27\%$ respectively. The testing accuracy, sensitivity, and specificity was $76.09\% \pm 11.24\%$, $95.16\% \pm 6.33\%$, and $73.09\% \pm 12.04\%$ respectively.

Testing set performance on Figure 3.2 shows 8 individual patient models with perfect specificity, meaning that for the times that the patients were in the transitioning state, the decision the model made matched perfectly. However, the specificity of these patients is not as high as in the training performance.

In Figure 3.3, there appeared to be an inverse correlation between length of

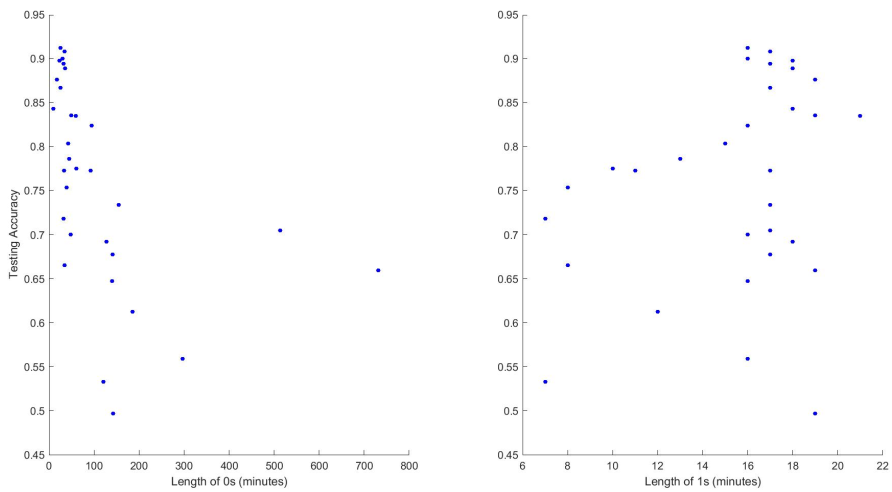


Figure 3.3: **Length of Data vs. Accuracy** Accuracy of testing set according to length of data in $\mathbf{y_i} = 0$ and $\mathbf{y_i} = 1$.

the $\mathbf{y_i} = 0$ state and accuracy of the models in the testing set. The Pearson correlation for length of time in $\mathbf{y_i} = 0$ vs accuracy was $\rho = -0.4553$. When omitting 2 cases of length of $\mathbf{y_i} = 0 > 400$ minutes, $\rho = -0.7474$. $\rho = 0.3332$ for the testing set correlation between $\mathbf{y_i} = 1$ and accuracy. The training set showed no such correlation.

3.2 Global Model

The global model utilized as many patients containing transitions as possible. Of the 137, 126 patients had data to sample $\mathbf{y_{i^1}} = 1$ from. There were 1851 i^0 patients available for age group and gender matching. The number of patients in each age group and gender are listed in Table 3.1.

The mean AUC of the ROC curve was 0.65 ranging from 0.61 to 0.70 across 50 iterations.

	i^1 (N=126)	i^0 (N=1851)
<1 mo	14	37
1-12 mo	30	279
12-24 mo	9	143
24-60 mo	28	367
60-144 mo	23	499
144-216 mo	22	526
Male	66	1019

Table 3.1: **Global Model Age and Gender**

The mean ± standard deviation for time point performance for i^0 patients, shown in Figure 3.5(a), was 63.9% ± 1.36% in accuracy, 91.08% ± 6.64% in sensitivity, and 21.87% ± 11.0% in specificity for the training set, and 11.95% ± 7.01%,92.67% ± 7.45%, and 11.91% ± 7.02% for the testing set. Figure 3.5(b) shows the training and testing performance for patient wide detection, with both training having 1.57% ± 2.86% accuracy and specificity, and 98.56% ± 1.99% accuracy and specificity in the testing set.

The mean ± standard deviation for time point performance in i^1 patients is shown in Figure 3.7(a). Training accuracy, sensitivity, and specificity were 63.9% ± 1.36%, 91.08% ± 6.64%, and 21.87% ± 11.0% respectively. Testing accuracy, sensitivity, and specificity were 13.68% ± 6.87%, 91.10% ± 6.62%, and 12.05% ± 7.07% respectively. Figure 3.7(b) shows patient wide detection statistics for i^1 patients. Accuracy and sensitivity were both 98.41% ± 2.19% in the training set, and 98.56% ± 1.99% in the testing set.

Using the method of normalizing each testing patient, the results were extremely close to normalizing across the entire testing set. The most substantial difference in performance was that the i^0 patients had an larger accuracy and specificity, and smaller sensitivity in time point classification at 13.82% ± 7.01%

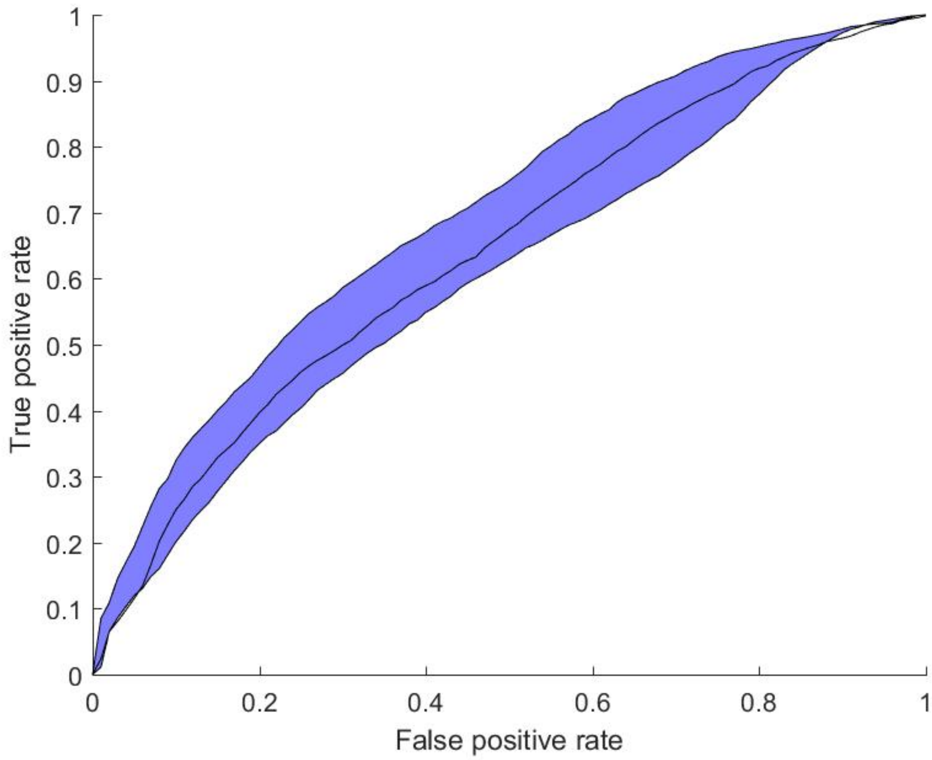
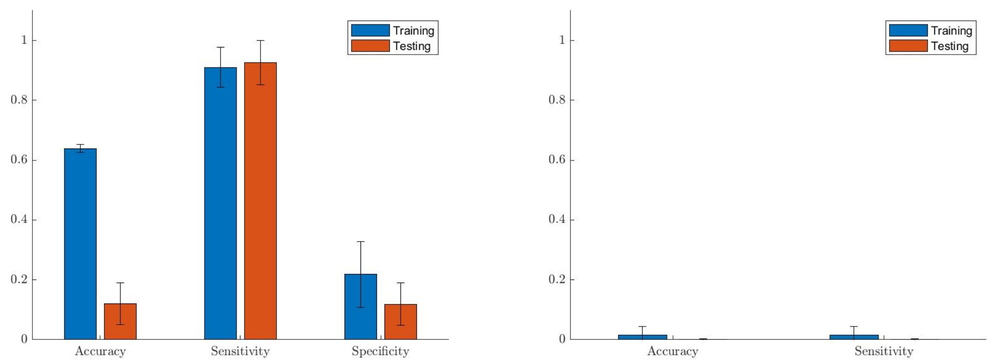


Figure 3.4: **ROC for Global Model** The black lines represent the curves with minimum, mean, and maximum, AUC across bootstrap iterations. The blue shaded region represents where the ROC curves of other iterations lay.

accuracy, $79.65\% \pm 12.81\%$ sensitivity, and $13.8\% \pm 7.02\%$ specificity. The rest of the exact performance statistics are listed in Appendix Table 5.1.



(a) Time Point Classification

(b) Patient-wide Detection

Figure 3.5: **Performance Statistics for i^0 Patients** Training and testing set accuracy, sensitivity, and specificity for patients who remain below transition threshold for the entire duration of PICU stay.

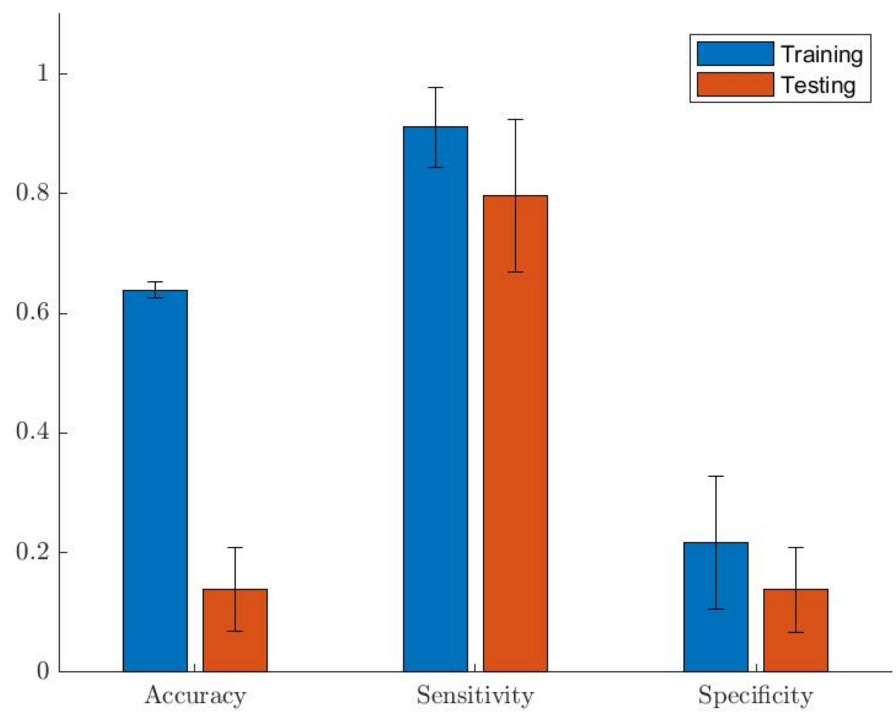
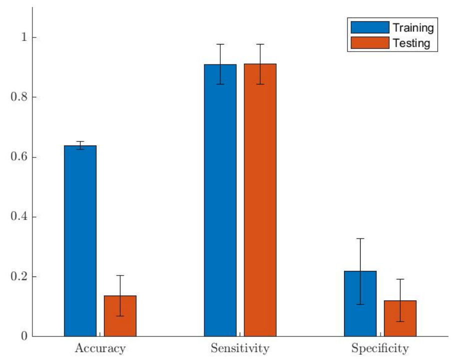
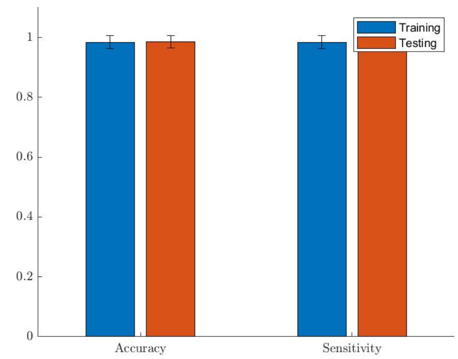


Figure 3.6: **Time Point Classification** using per patient normalization scheme on the testing set



(a) Time Point Classification



(b) Patient-wide Detection

Figure 3.7: **Performance Statistics for i^1 Patients** Training and testing set accuracy, sensitivity, and specificity for patients who meet transition criteria.

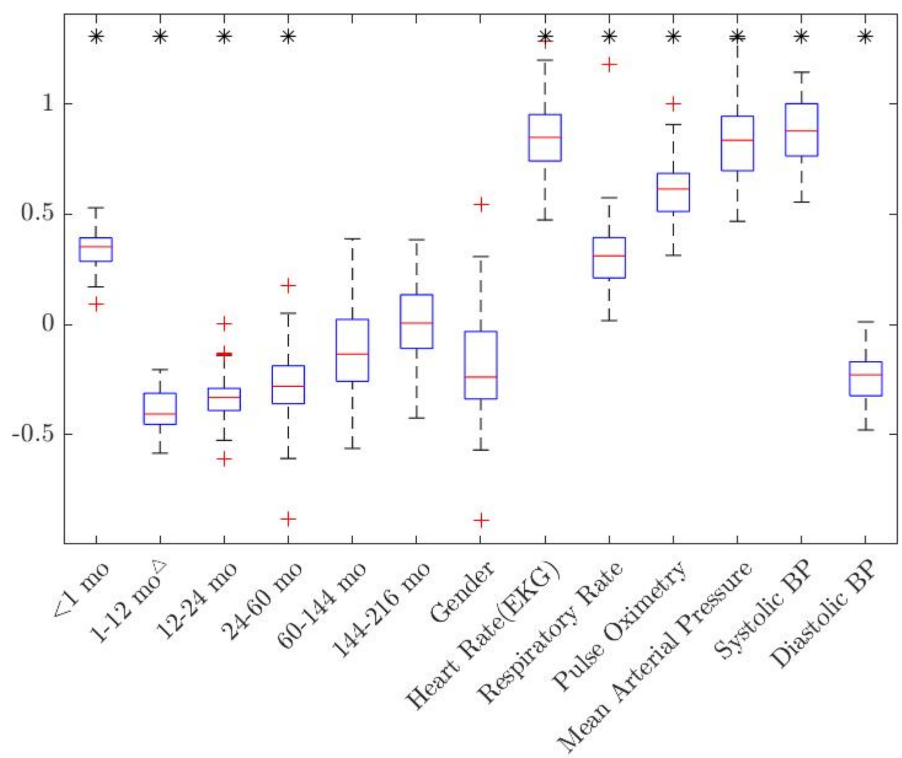


Figure 3.8: **Coefficient Values** Values for $\boldsymbol{\beta}$ obtained by the GLM. $^\triangle$ time limits are inclusive of the first number and strictly less than the second. *p$\leq$ .05

Chapter 4

Discussion and Conclusion

The complexity of MODS and its multiple and differing trajectories between patients in combination with the low prevalence of high severity samples makes computational model aided prediction a challenging task. This study provides a basis for further research on detecting the risk of patients who may transition into a severe state of multiple organ dysfunction. At the patient specific level, building GLM based computational models for severity detection shows promise, but improvements are necessary to make valid predictions on a more general population level.

4.1 GLM and Computational Modeling

The patient specific model performance suggests that GLM can be used to detect patients at risk of transitioning into severe states with organ failure at a highly overtrained level, but at a global population level, the model used in this study is lacking. In addition to the proposition mentioned above on including different patient groupings from commorbidity and length of stay, another point that could be improved upon is that criteria that was used to predict transitioning patients was a strict threshold, τ. From Figures 3.7 and 3.5, global model performance between patients who transition and patients who do not was similar at a time point level, but the patient wide detection showed a high tendency for the model to predict that patients are likely to transition. Many of the patients who were always below transition threshold were falsely detected as at risk using this strict threshold. Performance could possibly increase if a moving average threshold or cumulative risk score over a small window was implemented in the case of spikes in $\hat{\mathbf{y}}$, risk. However, at a cursory glance when plotting the $\hat{\mathbf{y}}$ across patients, many patients had longer periods where $\hat{\mathbf{y}}$ remains above τ.

Crone and Finlay [32] analyzed the effect of heterogeneous variations in large datasets on the performance of a classifying imbalanced classes using logistic regression, linear discriminant analysis, classification and regression trees, and neural networks. Three situations were considered: the over or under sampling method of balancing contribution from a binary classification problem, slowly incrementing class imbalance because the inherent population is imbalanced, and a combination of the sample size and balancing considerations. Logistic regression was found to have no benefit from artificially balancing the samples, and benefitted from using all data avialable [32].

Similarly, an experiment using ISOLET speech data, which contains various subjects enunciating letters of the alphabet from the University of California Irvine Machine Learning Repository, varied the prevalence of a target letter of the alphabet and compared regularized logistic regression, random forest, and soft-margin support vector machine model performance [33]. Accuracy, positive predictive value (PPV), sensitivity, and specificity were evaluated on each model with training prevelence varying from above 0 to 0.5. Accuracy, PPV, and specificity decreased, and sensitivity increased. The elements of speech recognition have much better precedence and model performance than the modeling of organ failure in this study. Therefore, seeing the change in performance of a well known logistic regression based classfier due to artificially increasing prevalence of the positive class suggests that the method of undersampling and class balancing in this study needs to be adjusted.

In line with this analysis of prevelence of positive class in the training data, the accuracy of the patient specific models decreased when patients with longer lengths of pre-transitioning data are artificially downsampled (Figure 3.3). Logistic regression finds the optimal set of coefficients which minimize deviance according to the distribution of the training data, and is the specific type of generalized linear model that uses logit link function used in this study. Thus, it makes sense that downsampling from the $\mathbf{y} = 0$ class led to a large sensitivity in the results and extremely low specificity. The number of patients who never transition outweighed the number of patients who do transition by nearly 16 times, and the window of transitioning time was a mere 2 hours compared to median length of stay at over 24 hours.

4.2 Clinically related factors and MODS Physiology

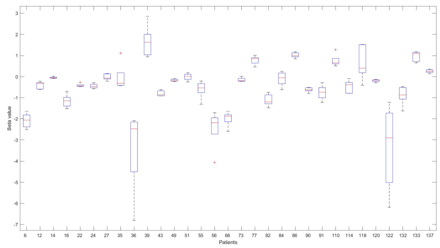
The models evaluated in this study used PELOD-2 to define the difference between a low risk and a patient at risk of severe organ dysfunction. However, the

ACCP/SCCM Consensus Conference Committee outlined two different pathways to MODS, where primary MODS occurs early, directly after an insult to the organs involved, and secondary MODS as a later response to some other inciting injury such as systemic inflamation response syndrome or sepsis [22]. The two pathways for MODS physiology suggest that it may be helpful to stratify patients into two groups depending on duration of PICU stay or time when transition occurs.

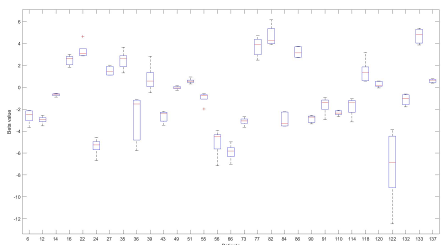
Proulx et al. [34] used the differentiation of primary MODS if patients had MODS at PICU admission, and secondary MODS if patients developed MODS during the first week after admission or later. Using this critera, Proulx et al. [1] and Tantaleán et al. [2] found the majority of the pediatric MODS patients as having primary MODS at 84.6% and 86% respectively. Similarly, the majority of patients in this study transitioned within 12 hours of PICU admission (the 3rd quartile for first transition time was 13.36 hours after admission).

Another consideration that could be further investigated are commorbidities; Bestati et al. [5] suggests that congenital cardiac disease may contribute to high hazard ratio related to cardiac dysfunction in neonates. It may be worth clustering patients by presence of commorbidities such as congenital diseases or surgical operations.

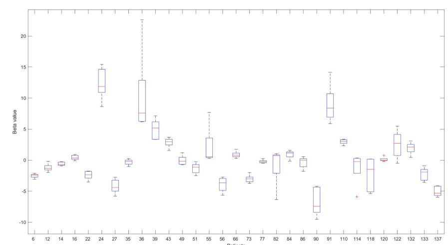
The $\boldsymbol{\beta}$ values in Figure 4.1 are trained from individual patients and too specific for comparison to the global population values, where interpretation of $\boldsymbol{\beta}$ is also dependent on age group and gender, but the variation between and within patients may reveal information with future further analysis on the above mentioned physiologically related factors and comorbidities.



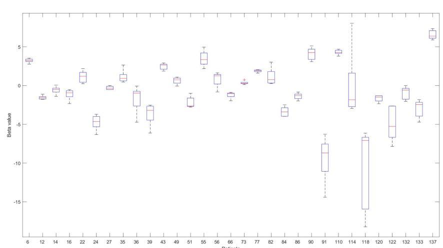
(a) Constant



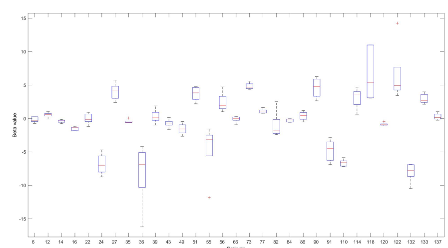
(b) Heart Rate



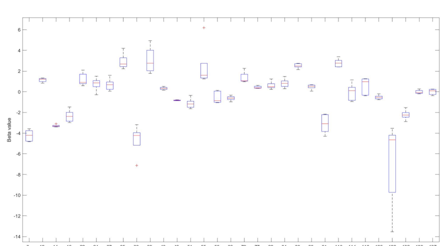
(c) Mean Arterial Pressure



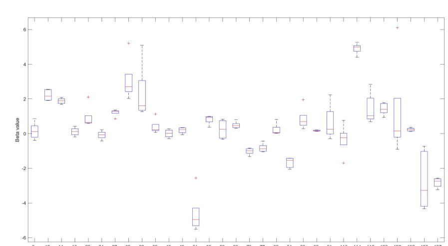
(d) Systolic Blood Pressure



(e) Diastolic Blood Pressure



(f) Respiration Rate



(g) Pulse Oximetry

Figure 4.1: **Patient Specific Coefficient Values** Values for $\boldsymbol{\beta}$ obtained by the GLM.

4.3 PELOD-2 Related Outcomes

The definition for severity defined in this study relied on the variables mean arterial pressure and presence of invasive ventilation that had high enough frequency to mark the transition with some confidence. This noticeably limited the types of patients used for modeling, and in the future, data sets with more regularly measured variables would be beneficial to early detection studies. In Table 2.2, the presence of individual organ PELOD-2 scores above zero during a desired transition are tallied. All of the selected patients who contain transitions have respiratory organ failure, and all but two of them have cardiovascular failure. This is expected from only relying on transitions caused by mean arterial pressure or invasive ventilation. There was also a fairly high number 119 (87%) of the 137 who had at least a score of 1 in neurological organ failure. Although, in the PELOD-2 score, GCS and pupillary reaction are the defining factors of neurologic organ dysfunction, GCS is the main source contributing to the score. The data obtained from EHR on pupil movement and dilation was highly sparse, with only a handful of patients from the entire 2512 patients containing information involving both pupils being simultaneously fixed and dilated.

The older PELOD score was found to be a significant prognostic factor for death [35], but the patients in the Leteurtre study were stratified into three categories of day 1 PELOD scores: $<$10, 10-19, and ≥ 20. Even the lowest score category included patients beyond what was considered severe in this study. An earlier model was built to classify the extreme transition in this study cohort, where selected patients were admitted below a PELOD-2 score of

5 and transitioned to a score ≥ 20. Perfect separation was achieved on the 2 patients under this criteria, without implementing the transitional window and purely classifying time points on training data. The definition of severity may use further exploration to achieve better model results, however the challenge to this is that more stringent transition definitions are likely to reduce the number of patients labeled with a transition.

Suggestions for further study include varying class balancing schemes to find the optimal method of sampling, testing other definitions of severity which incorporate other physiological factors of what is known about the progression of MODS, and varying the length of the transitional window. Calibration of these model definitions may lead to a model that is performs well across the general PICU population and can aid physician early diagnoses.

Bibliography

[1] Francois Proulx et al. "Timing and predictors of death in pediatric patients with multiple organ system failure." In: *Critical care medicine* 22.6 (1994), pp. 1025–1031.

[2] José A. Tantaleán et al. "Multiple organ dysfunction syndrome in children". In: *Pediatric Critical Care Medicine* 4.2 (2003), pp. 181–185. DOI: 10.1097/01.pcc.0000059421.13161.88. URL: https://doi.org/10.1097/01.pcc.0000059421.13161.88.

[3] AYT Goh, LCS Lum, and PWK Chan. "Brief report. Paediatric intensive care in Kuala Lumpur, Malaysia: a developing subspecialty". In: *Journal of tropical pediatrics* 45.6 (1999), pp. 362–364.

[4] Stéphane Leteurtre et al. "Validation of the paediatric logistic organ dysfunction (PELOD) score: prospective, observational, multicentre study". In: *The Lancet* 362.9379 (2003), pp. 192–197.

[5] Nawar Bestati et al. "Differences in organ dysfunctions between neonates and older children: a prospective, observational, multicenter study". In: *Critical Care* 14.6 (2010), R202. DOI: 10.1186/cc9323. URL: https://doi.org/10.1186/cc9323.

[6] Ana Lia Graciano et al. "The Pediatric Multiple Organ Dysfunction Score (P-MODS): development and validation of an objective scale to measure the severity of multiple organ dysfunction in critically ill children". In: *Critical care medicine* 33.7 (2005), pp. 1484–1491.

[7] Stéphane Leteurtre et al. "Development of a pediatric multiple organ dysfunction score: use of two strategies". In: *Medical Decision Making* 19.4 (1999), pp. 399–410.

[8] Stéphane Leteurtre et al. "PELOD-2". In: *Critical Care Medicine* 41.7 (2013), pp. 1761–1773. DOI: 10.1097/ccm.0b013e31828a2bbd. URL: https://doi.org/10.1097/ccm.0b013e31828a2bbd.

[9] William A Knaus et al. “APACHE II: a severity of disease classification system.” In: *Critical care medicine* 13.10 (1985), pp. 818–829.

[10] Jean-Louis Vincent and Rui Moreno. “Clinical review: scoring systems in the critically ill”. In: *Critical care* 14.2 (2010), p. 207.

[11] Francesca Innocenti et al. “Prognostic scores for early stratification of septic patients admitted to an emergency department-high dependency unit”. In: *European Journal of Emergency Medicine* 21.4 (2014), pp. 254–259. DOI: 10.1097/mej.0000000000000075. URL: https://doi.org/10.1097/mej.0000000000000075.

[12] Daliana Peres Bota et al. “The multiple organ dysfunction score (MODS) versus the sequential organ failure assessment (SOFA) score in outcome prediction”. In: *Intensive care medicine* 28.11 (2002), pp. 1619–1624.

[13] Andréanne Villeneuve et al. “Multiple organ dysfunction syndrome in critically ill children: clinical value of two lists of diagnostic criteria”. In: *Annals of Intensive Care* 6.1 (2016). DOI: 10.1186/s13613-016-0144-6. URL: https://doi.org/10.1186/s13613-016-0144-6.

[14] Martha C. Kutko et al. “Mortality rates in pediatric septic shock with and without multiple organ system failure”. In: *Pediatric Critical Care Medicine* 4.3 (2003), pp. 333–337. DOI: 10.1097/01.pcc.0000074266.10576.9b. URL: https://doi.org/10.1097/01.pcc.0000074266.10576.9b.

[15] Katri V. Typpo et al. “Day 1 multiple organ dysfunction syndrome is associated with poor functional outcome and mortality in the pediatric intensive care unit”. In: *Pediatric Critical Care Medicine* 10.5 (2009), pp. 562–570. DOI: 10.1097/pcc.0b013e3181a64be1. URL: https://doi.org/10.1097/pcc.0b013e3181a64be1.

[16] François Proulx et al. “The pediatric multiple organ dysfunction syndrome”. In: *Pediatric Critical Care Medicine* 10.1 (2009), pp. 12–22. DOI: 10.1097/pcc.0b013e31819370a9. URL: https://doi.org/10.1097/pcc.0b013e31819370a9.

[17] Michelle Ramírez. “Multiple organ dysfunction syndrome”. In: *Current problems in pediatric and adolescent health care* 43.10 (2013), pp. 273–277.

[18] Juan C. Jaramillo-Bustamante et al. "Epidemiology of sepsis in pediatric intensive care units". In: *Pediatric Critical Care Medicine* 13.5 (2012), pp. 501–508. DOI: 10.1097/pcc.0b013e31823c980f. URL: https://doi.org/10.1097/pcc.0b013e31823c980f.

[19] David F. Gaieski et al. "Impact of time to antibiotics on survival in patients with severe sepsis or septic shock in whom early goal-directed therapy was initiated in the emergency department". In: *Critical Care Medicine* 38.4 (2010), pp. 1045–1053. DOI: 10.1097/ccm.0b013e3181cc4824. URL: https://doi.org/10.1097/ccm.0b013e3181cc4824.

[20] Michael A. Puskarich et al. "Association between timing of antibiotic administration and mortality from septic shock in patients treated with a quantitative resuscitation protocol". In: *Critical Care Medicine* 39.9 (2011), pp. 2066–2071. DOI: 10.1097/ccm.0b013e31821e87ab. URL: https://doi.org/10.1097/ccm.0b013e31821e87ab.

[21] James D Wilkinson et al. "Mortality associated with multiple organ system failure and sepsis in pediatric intensive care unit". In: *The Journal of pediatrics* 111.3 (1987), pp. 324–328.

[22] American College of Chest Physicians and Society of Critical Care Medicine Consensus Conference Committee and others. "American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference: definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis". In: *Crit Care Med* 20 (1992), pp. 864–874.

[23] B Goldstein. "International Consensus Conference on Pediatric Sepsis. International pediatric sepsis consensus conference: definitions for sepsis and organ dysfunction in pediatrics". In: *Pediatr. Crit. Care Med.* 6 (2005), pp. 2–8.

[24] Nesrin O Ghanem-Zoubi et al. "Assessment of disease-severity scoring systems for patients with sepsis in general internal medicine departments". In: *Critical Care* 15.2 (2011), R95. DOI: 10.1186/cc10102. URL: https://doi.org/10.1186/cc10102.

[25] Jill S. Sweney et al. "Comparison of Severity of Illness Scores to Physician Clinical Judgment for Potential Use in Pediatric Critical Care Triage". In: *Disaster Medicine and Public Health Preparedness* 6.02 (2012), pp. 126–130. DOI: 10.1001/dmp.2012.17. URL: https://doi.org/10.1001/dmp.2012.17.

[26] Murray M Pollack, Kantilal M Patel, and Urs E Ruttimann. “PRISM III: an updated Pediatric Risk of Mortality score”. In: *Critical care medicine* 24.5 (1996), pp. 743–752.

[27] Nobuaki Shime et al. “Application of modified sequential organ failure assessment score in children after cardiac surgery”. In: *Journal of cardiothoracic and vascular anesthesia* 15.4 (2001), pp. 463–468.

[28] Micol Sandri et al. “Dynamic Bayesian Networks to predict sequences of organ failures in patients admitted to ICU”. In: *Journal of Biomedical Informatics* 48 (2014), pp. 106–113. DOI: 10.1016/j.jbi.2013.12.008. URL: https://doi.org/10.1016/j.jbi.2013.12.008.

[29] Chris Feudtner et al. “Pediatric complex chronic conditions classification system version 2: updated for ICD-10 and complex medical technology dependence and transplantation”. In: *BMC pediatrics* 14.1 (2014), p. 199.

[30] David A. Harrison et al. “A new risk prediction model for critical care: The Intensive Care National Audit & Research Centre (ICNARC) model”. In: *Critical Care Medicine* 35.4 (2007), pp. 1091–1098. DOI: 10.1097/01.ccm.0000259468.24532.44. URL: https://doi.org/10.1097/01.ccm.0000259468.24532.44.

[31] Peter McCullagh and John A Nelder. *Generalized linear models*. Vol. 37. CRC press, 1989.

[32] Sven F Crone and Steven Finlay. “Instance sampling in credit scoring: An empirical study of sample size and balancing”. In: *International Journal of Forecasting* 28.1 (2012), pp. 224–238.

[33] Nina Zumel and John Mount. *Does Balancing Classes Improve Classifier Performance?* 2015. URL: http://winvector.github.io/Prevalence/.

[34] Francois Proulx et al. “Epidemiology of sepsis and multiple organ dysfunction syndrome in children”. In: *Chest* 109.4 (1996), pp. 1033–1037.

[35] S. Leteurtre et al. “Daily estimation of the severity of multiple organ dysfunction syndrome in critically ill children”. In: *Canadian Medical Association Journal* 182.11 (2010), pp. 1181–1187. DOI: 10.1503/cmaj.081715. URL: https://doi.org/10.1503/cmaj.081715.

Chapter 5

Appendix

5.1 Figures and Tables

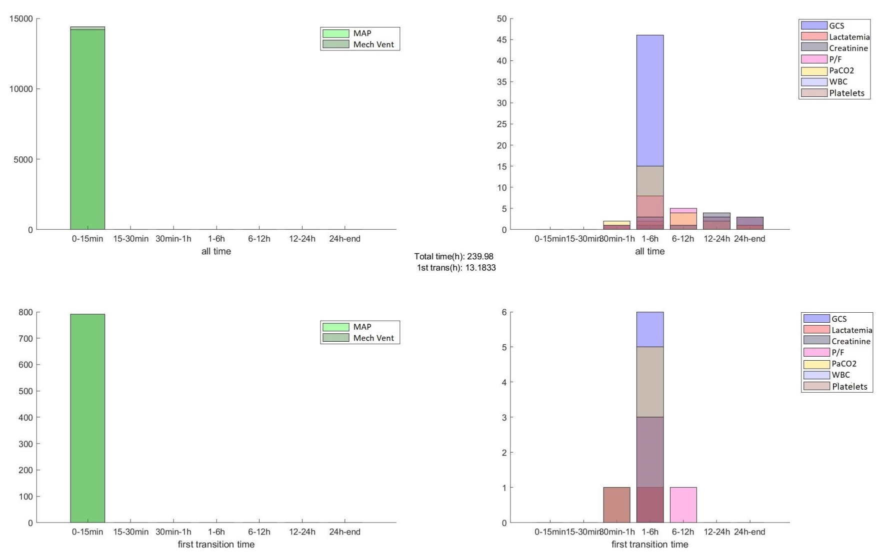


Figure 5.1: **Frequency of Physiological Variables** Example of variable frequency occuring in one patient. The x-axis represents difference in time from previous data sample or start of data collection. The y-axis represents histogram count of how many occurences this interval is present. The top two quadrants shows entire patient admission time and the bottom shows up to first transition time only. Opacity of the bars is lowered to show that bars are overlapped, not stacked.

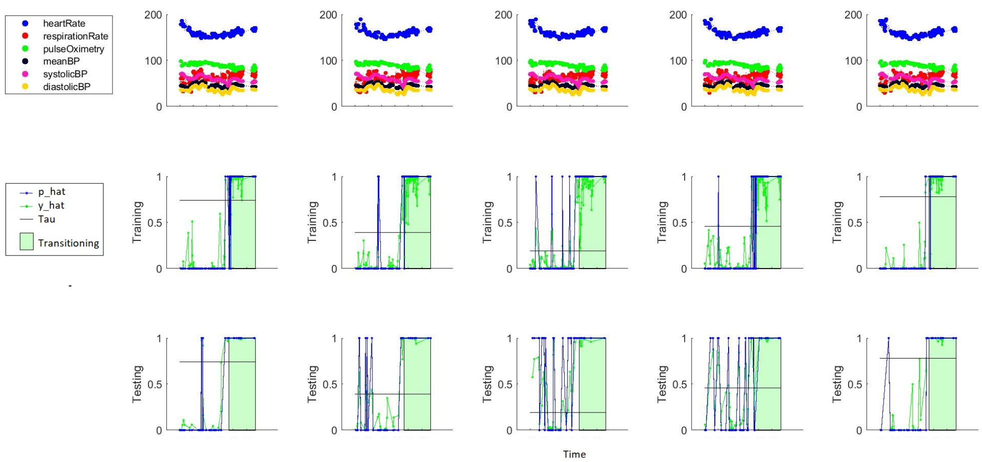


Figure 5.2: **Example GLM iterations** of the patient specific models.

		Time Point Classification			Patient Detection		
		Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
i^1 Patients	Training	63.91% ± 1.36%	91.15% ± 6.69%	21.73% ± 11.09%	98.39% ± 2.18%	98.39% ± 2.18%	-
	Testing	15.47% ± 6.89%	90.92% ± 6.70%	13.9% ± 7.07%	98.52% ± 1.97%	98.52% ± 1.97%	-
i^0 Patients	Training	63.91% ± 1.36%	91.15% ± 6.69%	21.73% ± 11.09%	1.57% ± 2.86%	-	1.57% ± 2.86%
	Testing	13.82% ± 7.01%	79.65% ± 12.81%	13.8% ± 7.02%	0.2% ± 0.08%	-	0.2% ± 0.08%

Table 5.1: **Global Model Performance Statistics** using per patient normalization on the testing set

5.2 Data Processing

5.2.1 PELOD-2 Variables

Glasgow Coma Score values were invalidated during all time points between recorded administrations of one unique sedative. The list of sedatives was: Dexmedetomidine, Fentanyl, Hydromorphone, Ketamine, Midazolam, Morphine, Propofol, Remifentanil.

Pupils were only marked as fixed and dilated (needed for the Leteurtre et al. [8] PELOD-2 definition) if both pupils contained information indicating that, within a time interval up to and including 5 minutes, both pupils were ≥ 3 mm as well as non-reactive or fixed. In logic notation, documentation of each pupil within 5 minutes must say:

$$\text{left pupil: } \geq 3 \text{ mm AND (fixed or non-reactive)}$$

$$\text{AND}$$

$$\text{right pupil: } \geq 3\text{mm AND (fixed or non-reactive)}$$

for pupils to contribute the PELOD-2 score. Time points with indications of chemical dilation by Opthamology were discarded.

Mean arterial pressure was taken from non-invasive measurements first, and then arterial measurements if non-invasive measurements were missing. Additionally, arterial MAP was counted as missing if values exceeded 120 mmhg.

PaO_2/FiO_2 ratio was calculated from the measurements within 1 hour of each other, order of measurement was disregarded. In the case that more than one measurement existed within this acceptable window, PaO_2/FiO_2 was calculated using the values that would give the most severe value to the ratio. The time of the first measured value in the ratio was assigned to the PaO_2/FiO_2 ratio.

Mechanical or invasive ventilation was identified by a list of strings recorded in EHR flowsheet data. After positively identifying a patient on mechanical ventilation, the status was held for every minute up to an hour or until the next identifier occurred. Non-invasive ventilation was also identified, so times where both methods were supposedly present were removed for any mode of ventilation due to the uncertainty.

The PELOD-2 score was calculated following Leteurtre et al. [8]. A sliding window of 24 hours was implemented, and the PELOD-2 score was assigned based on the sum of scores determined using the worst values taken for each variable within the previous 24 hours. The 24 hour sliding windows began with one timepoint at the start of patient data availability, and expanded up until the full 24 hours so a PELOD-2 score existed for all times when data existed for a patient.The time of the score was assigned to the last time point existing within a window. If no data were available for a particular variable within the prior 24 hours, a score of 0 (normal) was assigned for that variable. If a variable was measured more than once in the prior 24 hours, the worst value was used in calculating the score.

MICHELLE CHYN
49 Norgate Rd, Manhasset, NY 11030, (516)507-7178, mchyn1104@gmail.com

EDUCATION

Johns Hopkins University, Baltimore MD
Master's of Science and Engineering — October 2018
Bachelor of Science in Biomedical Engineering — Minor: Computational Medicine — May 2016
Awards: Dean's List Award for Academic Excellence Spring 2013-Spring 2015, Schrodel Endowed Scholarship Recipient December 2014 and December 2015

PROJECT EXPERIENCE

Master's Research — **Johns Hopkins University**
November 2016-Present

- Import and process large string and numeric data sets from electronic health record extracts in MATLAB
- Used server cluster computations to perform modeling such as general linear modeling
- Present and discuss findings and next plan of action with professors and clinicians weekly

Undergraduate Research — **Johns Hopkins University**
January 2015-May 2015

- Used network graph based analysis on EEG data in MATLAB to find statistical significance in regional cortical activity during a decision making gambling task
- Wrote and submitted 4 page paper to IEEE Engineering in Medicine and Biology Society Conference

Introduction to Computational Medicine Class — **Johns Hopkins University**
September 2015-December 2015

- Completed 6 projects with varying data collection and modeling techniques in one school semester with a team of 5 students mimicking cutting edge biomedical research
- Wrote reports and presented findings to the class for each project within a time limit of 7 minutes

BME Senior Design — **Johns Hopkins University**
September 2014 – May 2015

- Independently learned circuit design to implement wireless Electrooculography (EOG)
- Adjusted specifications of existing Bluetooth technology to incorporate in wireless EOG in a team of 3

WORK EXPERIENCE

Study Consultant — **Johns Hopkins Center for Academic Advising**
October 2014-Present

- Advised undergraduate and graduate students on methods of studying for classes of varying topics
- Taught students learn how to improve time management skills and to balance social and academic life during high stress situations
- Led workshops for training new and current study consultants

Systems Bioengineering Lab Teaching Assistant — **Johns Hopkins**
September 2016-May 2017

- Prepare materials for and lead a 4 hour lab section with 25 students biweekly
- Held office hours and review sessions, graded lab reports and exams

Physics Undergraduate Teaching Assistant — **Johns Hopkins**
January 2014-May 2014

- Supervise students through Physics Electricity and Magnetism problems sets in section
- Worked individually with 20 college level students weekly

Relevant Course Work: Audio Signals Processing, Systems Bioengineering I, II, and III, Statistical Connectomics, Medical Imaging Systems, Signals, Systems, and Controls, Circuits, Networks, Learning Theory, Models of the Neuron, Introduction to Statistics, Machine Learning, Representations of Choice
Programming: MATLAB, R, familiarity with Java, Python, LaTeX, TORQUE, GitHub